Dual Modality Prompt Tuning for Vision-Language Pre-Trained Model

Yinghui Xing*, Qirui Wu*, De Cheng✉, Shizhou Zhang, Guoqiang Liang, Peng Wang, Yanning Zhang.

***Abstract*—With the emergence of large pretrained vison-language models such as CLIP , transferable representations can be adapted to a wide range of downstream tasks via prompt tuning. Prompt tuning probes for beneficial information for downstream tasks from the general knowledge stored in the pretrained model. A recently proposed method named Context Optimization (CoOp) introduces a set of learnable vectors as text prompts from the language side. However, tuning the text prompt alone can only adjust the synthesized "classifier", while the computed visual features of the image encoder cannot be affected, thus leading to suboptimal solutions. In this paper, we propose a novel dual-modality prompt tuning (DPT) paradigm through learning text and visual prompts simultaneously. To make the final image feature concentrate more on the target visual concept, a class-aware visual prompt tuning (CAVPT) scheme is further proposed in our DPT. In this scheme, the class-aware visual prompt is generated dynamically by performing the cross attention between text prompt features and image patch token embeddings to encode both the downstream task-related information and visual instance information. Extensive experimental results on 11 datasets demonstrate the effectiveness and generalization ability of the proposed method. Our code is available in https://github.com/fanrena/DPT.**

***Index Terms*—Few-shot learning, Transfer learning, Image Classification, Prompt Tuning, Vision-Language Model**

I. Introduction

Recently, studies in large-scale vision-language models (VLM), such as CLIP [1] and ALIGN [2], have achieved remarkable progress in representation learning [3]–[5]. Benefiting from huge amounts of image-text data, the pretrained large-scale vision-language model is able to learn open-set visual concepts generated from natural language, thus further allowing zero-shot transfer to downstream tasks. Specifically, the vision-language model is composed of two components: the image encoder and the text encoder. When a new classification task arrives, one can synthesize the classifier by feeding the natural language description of the classes to the text encoder. Then, the similarity between the "classifier" and the image features generated by the image encoder is computed.

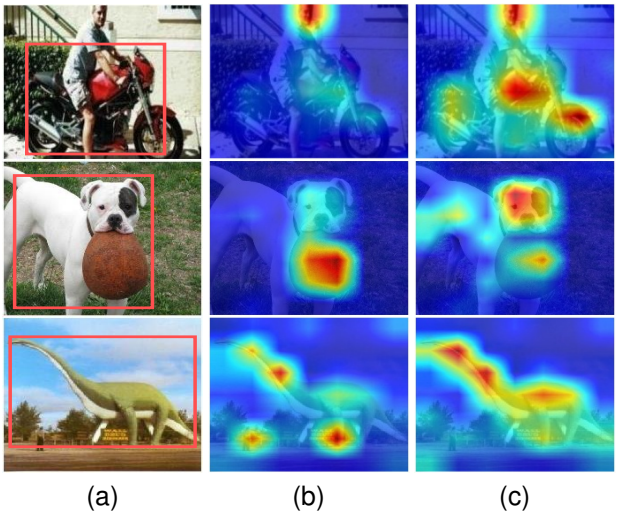

Fig. 1. Visualization of the attention map of the image encoder. (a) Original Image. (b) Zero-Shot CLIP/CoOp. (c) Our DPT. The images are selected from OxfordPets and Caltech101. The GT annotated object is marked by a red box. Best viewed in color.

However, adapting these pretrained large-scale vision-language models efficiently to downstream tasks demonstrates its own challenge. Recent studies show that "prompting" is a simple and effective method [1], while designing a proper prompt is a nontrivial task. It always requires extensive domain expertise and takes a significant amount of time for manual word tuning. Usually, even with massive tuning, we cannot guarantee that the obtained prompt is optimal for downstream tasks.

Recent studies on prompt learning for vision representation have been mainly inspired by some prompt tuning approaches in natural language processing (NLP) [6]–[8], e.g., the representative CoOp [9]. These methods proposed modeling learnable contexts in prompt using continuous representations and then trained the model with these learnable prompts in an end-to-end way while keeping the pretrained parameters fixed. Although these methods have achieved great success and show promising performance, they only learn prompts for the text encoder.

From the perspective of conventional visual recognition, a typical vision model can be roughly divided into a feature extraction module and a classifier. Similarly, the process of feeding the text prompt into the text encoder can be viewed as the synthesis of a classifier, and the image encoder that extracts

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62101453, Grant 62201467, and Grant 62176198; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110544; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022JQ-686, 2019JQ-158, and in part by the Project funded by China Postdoctoral Science Foundation under Grant 2022TQ0260, and in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313088.

Yinghui Xing, Qirui Wu, Shizhou Zhang, Guoqiang Liang, Peng Wang, Yanning Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi′an, China. Yinghui Xing is also with the Research & Development Institute of Northwestern Polytechnical University in Shenzhen. De Cheng is with School of Telecommunications Engineering, Xidian University, Xi′an, China. Corresponding author is De Cheng (email: dcheng@xidian.edu.cn)

*The first two authors equally contributed to this work.

the visual features. Assume that the large-scale pretrained vision-language models have already captured most of the general knowledge (visual concepts) for the downstream tasks. What the prompting mechanism does is to query the suitable information, which is beneficial to the downstream tasks, from the pretrained model. As shown in Figure 1, for an input image with multiple visual objects (concepts), e.g., the first case contains a person and a motorbike, the image encoder extracts all the visual features of the objects, i.e., the attention maps of Zero-Shot CLIP and CoOp highlight both the person and motorbike. However, the downstream task requires the output class label to be "motorbike"—the ground-truth annotation. CoOp tries to enable the model to output "motorbike" by adjusting the "classifier" alone while keeping the given highlighted "person" and "motorbike" visual features unchanged. There is a consensus in the vision community that features matter [10]! Therefore, we believe that adopting prompt tuning for the text encoder alone while directly utilizing the fixed image encoder for the downstream task is suboptimal. In this paper, we introduce visual prompts in the image input space and propose a dual-modality prompt tuning (DPT) paradigm by learning text prompts and visual prompts for both the text and image encoder simultaneously thus aiming at adapting the pretrained model to downstream tasks by adjusting both the "classifier" and "visual features".

Specifically, for visual prompt tuning in a ViT-based image encoder, we introduce a small number of trainable parameters in the input of the transformer blocks while keeping the pretrained image encoder fixed. Inserting visual prompts can directly adjust the image patch token embeddings. image features, through the self-attention weights and absorbing the prompt-derived value vectors. To make the pretrained model better transfer to the downstream task, we further introduce a class-aware visual prompt tuning (CAVPT) mechanism into our DPT framework to help the final obtained image feature concentrate more on the target visual concept. Thus we aim at encoding both the task-related information and visual instance information into the visual prompts, The class-aware visual prompt is dynamically generated by performing cross attention between text prompt features and visual image patch embeddings and is expected to include richer semantic features of the target visual objects. Thus, the final obtained image feature, which is computed by absorbing the information from the image patch embeddings and our class-ware visual prompts, can concentrate more on the classes corresponding to the downstream tasks. Finally, the proposed overall DPT paradigm is learned with text prompts, visual prompts, and class-aware visual prompts simultaneously. As shown in Figure 1, tuning the pretrained models with our DPT shows a more focused task-aware visual attention area.

The main contributions of this paper can be summarized in terms of the following three aspects:

- The proposed method demonstrates a new dual-modality prompt tuning paradigm for tuning the large pretrained vision-language model by simultaneously learning the visual and text prompts from the ends of both the text and image encoders.
- To encourage the visual prompts to explicitly contain downstream task-related information, we further introduce the class-aware visual prompt into our DPT. It is dynamically generated by performing cross attention between text prompt features and visual token embeddings.
- Extensive experimental results on 11 datasets demonstrate the effectiveness of the proposed method and shows its superiority to other prompt-tuning approaches by a large margin, as well as its generalization ability.

The remainder of this paper is organized as follows. Section II introduces the related works. Details of our proposed method are elaborated in Section III. In Section IV, we report the results of comprehensive experiments on 11 datasets used in prompt tuning, which demonstrates the effectiveness of our method. Finally, the conclusion of our work is presented in Section V.

II. Related Work

A. Vision-Language Pretrained Models

Learning visual representations under the supervision of natural language has been demonstrated to be effective and has attracted much attention [1], [2], [11], [12]. For vision-language models, image-text matching and cross-modal contrastive learning are two important issues. In CLIP [1], two encoders related to the vision and language modalities are designed, and these image and text embeddings are then aligned using a symmetric cross entropy loss. Similarly, ALIGN [2] also utilizes a dual-encoder architecture, but it projects the image and text embeddings to the same semantic space to calculate the similarity scores between vision and language modalities This makes the vision-language interaction more efficient. Both these models are pretrained on large-scale image-text datasets with the contrastive loss, and can be transferred to downstream tasks. Research on transferring CLIP to various downstream tasks, such as image classification [9], [13]–[15], video-text retrieval [16], tracking [17], and so on [18]–[21] is thriving. To boost the performance of CLIP to downstream tasks, CLIP-Adapter [13] introduced feature adapters on either visual or language branches and fine-tuned them on the few-shot classification task. Zhang et al. [14] further proposed a training-free CLIP-Adapter (*i.e.,* TIP-Adapter), which creates the weights by a key-value cache model constructed from the few-shot training set. With much less training, TIP-Adapter is more efficient than CLIP-Adapter. As an alternative framework to reduce the gap between objective forms of model pretraining and fine-tuning, prompt-based learning has become an active topic in both NLP and computer vision communities. However, the discrepancy between the two different modalities causes difficulties in tuning the prompt. Recently, Zhou *et al.* [9] proposed a context optimization (CoOp) strategy to automatically learn the optimal prompts, which greatly boosts the recognition accuracy. Our work also focuses on transferring the pretrained vision-language model to downstream tasks through prompting.

B. Prompt Learning

Prompt learning originated from the NLP community [6], [7], [22] and originally referred to the application of a fixed

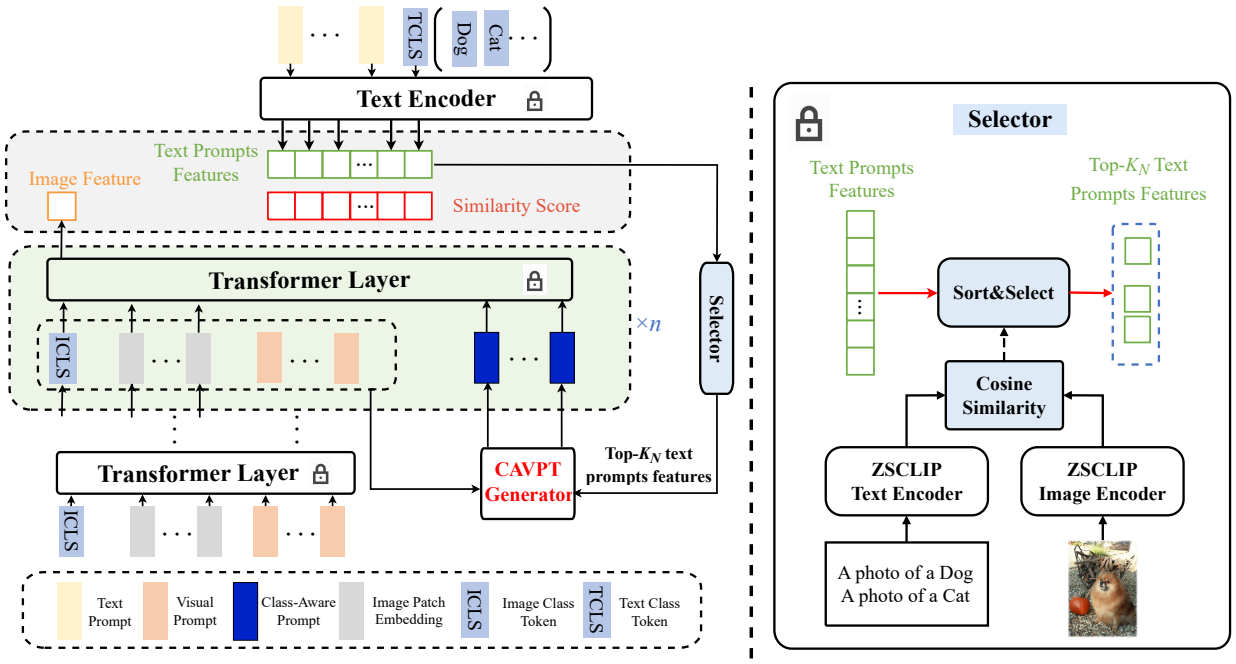

Fig. 2. The overall architecture of our proposed DPT method. It consists of three learnable components: text prompt, visual prompt and class-aware visual prompt generated from a Class-Aware Visual Prompt Tuning (CAVPT) generator module whose detailed architecture is illustrated in Fig. 3.

function to the input tokens, which provides an instruction about the task to the model. In the computer vision community, prompt learning has been explored in both visual models [23]–[25] and vision-language models [1], [9], [15], [18], [26]. In particular, visual prompt tuning (VPT) [23] has achieved significant performance gains with only a small amount of additional parameters, *i.e.*, prompts, in the input space. Vision-language models have been investigated in image classification [9], [15], [26]–[29], video recognition [30], and cross-modal learning [31]–[33]. Among them, CoOp [9] achieves continuous prompt optimization from downstream data to adapt the pretrained vision-language models. However, CoOp may introduce improper prompt tuning steps, which could hamper general knowledge probing [26]. To improve the generalization ability of CLIP, Zhu *et al.* [26] proposed a novel prompt tuning method, namely, *i.e.,* ProGrad, to address the conflicts between each tuning step and the general knowledge CLIP has predicted. Conditional CoOp (CoCoOp) [15] extended CoOp by learning an input-conditional token for each image to improve the cross-domain generalization ability of CoOp. Motivated by the fact that contrastive loss can improve the generalization ability of models, Sahoo *et al.* [34] introduced a contrastive prompt tuning approach. It augmented the standard cross-entropy loss with two additional contrastive loss terms to learn generalizable prompts without introducing any additional parameters. Lu *et al.* [27] learned the output embeddings of prompts instead of the input embeddings and employed a Gaussian distribution to model them effectively. Bahng *et al.* [28] proposed a prompting method for CNN networks to adapt the pretrained vision-language models to downstream tasks. In contrast, Zhang *et al.* [29] used a neural architecture search algorithm to identify the optimal configuration with adapters and prompts as small components.

Most of the existing methods tune the prompts in the text encoders alone and neglect the clues in visual features. Our work proposes a dual-modality prompt tuning paradigm, which introduces both the text prompt and visual prompt for the vision-language model. Furthermore, a class-aware visual prompt is proposed to enable the image feature to pay more attention to the target foreground object for downstream tasks.

C. Transfer Learning

Benefiting from the large scale of annotated data, the performance of deep neural networks has been greatly boosted. However, due to labeling costs, the collection of large-scale training datasets with accurate annotations is cumbersome [14]. Transfer learning [35]–[39] that aims to transfer general knowledge from one domain to some related domains with limited training data, has been proven to be a possible solution to few-shot learning [40]–[47]. Some works have tried to tune a small number of parameters while keeping most of the parameters of pretrained models frozen. For example, [37] adapted the pretrained network by training a lightweight side network that was fused with the frozen pretrained network via summation. [38] proposed a new memory-efficient bias module, *i.e.* the lite residual module, to refine the feature extractor by learning small residual feature maps. Rebuffi *et al.* [39] introduced a residual adapter to the model and only trained the adapter network to improve the accuracy of domain-specific representations.

On the other hand, some self-supervised learning-based methods, such as MoCo [48], BYOL [49], and MAE [50], can also alleviate the requirement of large-scale training data. Recently, vision-language models pretrained on large-scale image-text pairs have demonstrated their superiority. Therefore, it is crucial to excavate the potential of these models for downstream tasks. This paper focuses on transferring

knowledge learned from them to downstream tasks through prompting.

III. Methodology

In this section, we first revisit the CLIP model. Then, we elaborate each component of the proposed dual-modality prompt-tuning (DPT) paradigm, including text prompts, visual prompts and class-aware visual prompts. The framework of our proposed DPT is illustrated in Figure 2. Finally, we provide the loss function of DPT and a warm-up strategy to accelerate the training process.

A. Contrastive Language-Image Pretraining (CLIP) Model

The CLIP model aims to align the image feature space and text feature space, which enables the model to have the capability of zero-shot transfer to downstream tasks. CLIP is composed of two encoders: one is designed for images, and the other is designed for text. The text encoder adopts a transformer [51] to encode the text information The image encoder can either be a CNN model, such as ResNet [5], or a vision transformer, such as ViT [52]. In our method, we choose ViT as the image encoder to be compatibile with the visual prompt in [23].

With a tremendous number of 400 million pairs of image-text samples, CLIP is trained under the contrastive learning framework, where the associated image and text are treated as positive samples, while the non-associated samples are treated as negative samples. After that, all the parameters of the pretrained CLIP model are kept frozen for downstream tasks without any fine-tuning. In downstream tasks, a hand-crafted prompt is fed into the text end to synthesize a zero-shot linear classifier by embedding the class names of the target dataset. Taking the classification task as an example, the "[CLASS]" token can be first extended by a template, such as "a photo of a [CLASS]". Then, the sentence is treated as a prompt and is encoded by the text encoder to derive a weight vector $\mathbf{w}_i$, $i = \{1, ..., K\}$, where K is the total number of categories. At the same time, image features $\mathbf{x}$ are obtained by the image encoder. The prediction probability can be calculated by

$$p(y = i \mid \mathbf{x}) = \frac{\exp\left(\text{sim}\left(\mathbf{x}, \mathbf{w}_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}\left(\mathbf{x}, \mathbf{w}_j\right)/\tau\right)}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ represents the computation of cosine similarity, and τ is the temperature coefficient learned by CLIP.

B. Text Prompt and Visual Prompt

Text Prompt. It is known that hand-crafted prompts for the CLIP model may take considerable time and require expertise for word tuning, as a slight change in wording may lead to significant performance degradation. Motivated by prompt tuning in NLP models, CoOp [9] introduced a set of tunable word embedding vectors to learn machine-favorable prompts for the text end, which we call text prompts. Let $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M\}$ denote M learnable context vectors, and the word embedding of the text class token be represented by $\mathbf{c}_i$, $i = \{1, ..., K\}$; then, the prompt for the i_{th} class

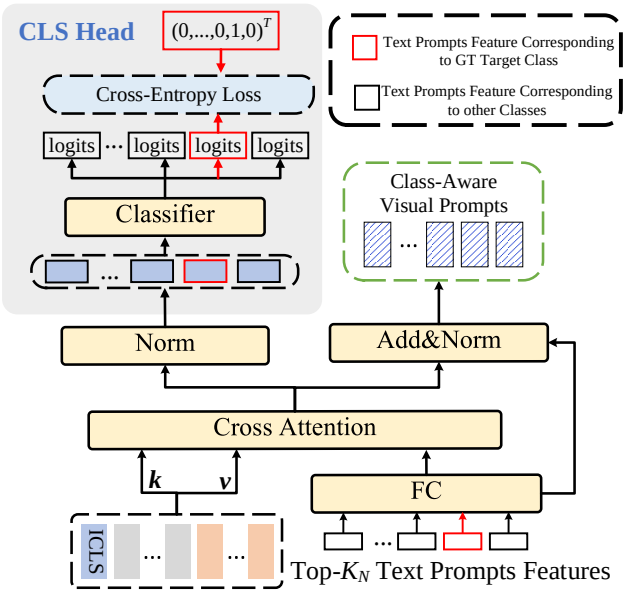

Fig. 3. The detailed architecture of the proposed class-aware visual prompt tuning (CAVPT) generator module.

can be denoted as $\mathbf{t}_i = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M, \mathbf{c}_i\}$. By forwarding $\mathbf{t}_i$ into the text encoder $g(\cdot)$, we can obtain a classification weight vector for the i_{th} visual concepts. The corresponding prediction probability can be calculated by

$$p(y = i \mid \mathbf{x}) = \frac{\exp\left(\text{sim}\left(\mathbf{x}, g(\mathbf{t}_i)\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}\left(\mathbf{x}, g(\mathbf{t}_j)\right)/\tau\right)}, \tag{2}$$

where $\mathbf{x}$ represents the extracted image features, and $g(\cdot)$ denotes the text encoder.

Visual Prompt. For vision-language models, there are two encoders for visual and language modalities. Tuning text prompts alone is not enough to reduce the gap between pretrained and downstream tasks, thus leading to suboptimal results. Motivated by the visual prompt tuning (VPT) [23] proposed for tuning vision transformers, we introduce a visual prompt into the image encoder of the CLIP model. The image patches $\left\{\mathbf{I}_j \in \mathbb{R}^{3\times h\times w} \mid j \in \mathbb{N}, 1 \le j \le N_p\right\}$ are first embedded into a d-dimensional latent space as follows:

$$\mathbf{e}_0^j = \text{Embed}\left(\mathbf{I}_j\right) \quad \mathbf{e}_0^j \in \mathbb{R}^d, j = 1, 2, \ldots N_p. \tag{3}$$

Let $\mathbf{E}_l = \left\{\mathbf{e}_l^j \in \mathbb{R}^d \mid j \in \mathbb{N}, 1 \le j \le N_p\right\}$ and $\mathbf{P}_l = \left\{\mathbf{p}_l^i \in \mathbb{R}^d \mid i \in \mathbb{N}, 1 \le i \le P\right\}$ represent a collection of image patch embeddings and visual prompts for the l_{th} transformer layer, respectively. Suppose $\mathbf{s}_l \in \mathbb{R}^d$ is a learnable class token in the image encoder, which is different from the text class token used in text prompt that the latter is a category-related word embedding. There are two versions of visual prompts, VPT-Shallow and VPT-Deep, in [23]. We empirically found that VPT-Deep can achieve superior performances (see Table I), and hence we take VPT-Deep into our implementation in Section IV.

Visual prompts are introduced to each of the transformer layers, that is,

$$[\mathbf{s}_l, \_, \mathbf{E}_l] = \Phi_l\left([\mathbf{s}_{l-1}, \mathbf{P}_{l-1}, \mathbf{E}_{l-1}]\right), l = 1, 2, \ldots, L. \tag{4}$$

TABLE I
MAIN RESULTS OF 11 DATASETS UNDER 16-SHOTS SETTING.

Methods	EuroSAT	Caltech101	Oxford Flowers	Food101	FGVC Aircraft	DTD	OxfordPets	Stanford Cars	Sun397	UCF101	ImageNet	Average
ZSCLIP [1]	45.49	91.28	66.63	80.62	19.08	44.03	87.38	60.19	62.06	63.52	59.61	61.81
CoOp [9]	83.12	94.45	95.07	78.20	33.94	67.20	88.88	75.79	72.31	79.10	66.55	75.87
CoCoOp [15]	74.99	94.01	79.97	**82.36**	23.64	59.34	90.98	64.25	69.75	73.13	65.07	70.68
ProGrad [26]	82.49	95.18	94.60	81.15	32.50	65.98	90.43	74.85	73.22	78.52	66.60	75.96
ProDA [27]	83.28	95.5	95.98	81.89	34.68	**70.76**	90.6	77.64	**75.07**	81.85	**67.62**	77.72
VPT	**92.17**	94.85	93.80	81.29	39.98	67.16	90.32	72.03	69.84	80.17	64.17	76.89
VLP	91.90	95.10	96.05	78.42	42.92	68.06	90.33	78.81	72.12	**82.04**	66.91	78.42
DPT	91.16	**95.61**	**96.60**	79.25	**48.37**	70.16	**91.22**	**82.55**	70.97	81.43	66.85	**79.47**

Generally, performance is positively correlated with prompt depth. Therefore, we utilize VPT-Deep in our model. $\mathbf{s}_L$ is then projected by a linear projection layer LP to obtain the final image feature. For simplicity, the whole process of image feature extraction can be represented by

$$\mathbf{x}' = f([\mathbf{s}_0, \mathbf{P}_0, \cdots, \mathbf{P}_L, \mathbf{E}_0]), \tag{5}$$

where $f(\cdot)$ denotes the image encoder.

Note that the calculation process of the image encoder, *i.e.*, the ViT model, can be viewed as a process of global scene reasoning, and $\mathbf{s}_l$ pools the visual concepts from the image patch embeddings layer-by-layer. With the help of visual prompts, the target visual concept corresponding to the downstream task may be further highlighted in $\mathbf{s}_l$ via the self-attention operation in each transformer layer. By inserting visual prompts into each transformer layer, the self-attention operation for $\mathbf{s}_l$ can be affected in two ways, as both the keys and values are prepended through visual prompts: 1) The attention weights can be affected to allow $\mathbf{s}_l$ to concentrate more on the image patch embeddings, which includes the target concept; 2) The visual prompts also serve as value vectors for the self-attention operation and thus $\mathbf{s}_l$ may absorb additional information that visual prompts learned.

However, naive visual prompts are devised as unconstrained learnable vectors, and they can only learn some downstream task-related information implicitly by tuning the prompts on downstream task datasets. In this work, we propose class-aware visual prompt tuning (CAVPT) to generate visual prompts by utilizing both task-related information from the text side and instance wise information from the visual side.

C. Class-Aware Visual Prompt Tuning

Class-aware visual prompts aim to explicitly encode task-related information. Our CAVPT generator takes two sides of inputs, the instance-wise information from the visual side and the task-related information from the text side. The text prompts features computed by the text encoder with *all the text class tokens* well represents *the task-related information*. However, when we input the text prompts features with all the text class tokens into the CAVPT generator, the computational complexity of CAVPT generator is linearly increased with the number of classes on each downstream task. To reduce the computational complexity of our CAVPT generator into constant, we select top-K_N text prompts features with the help of a Zero-Shot CLIP Inference module (the right part of Figure 2). Note that the final performance is not sensitive to K_N and The final performance fluctuates with only $0.1\% \sim 0.2\%$ when setting different K_N. Then, feeding the text prompts with the top-K_N text class token [CLASS] into the text encoder produces K_N feature vectors, *i.e.*, $\mathbf{g}_j \in \mathbb{R}^D, 1 \leq j \leq K_N$, in which the task-related information are encoded. A class-aware visual prompt is generated dynamically by performing cross-attention between text prompt features from the text side and the inputs of the transformer block from the visual side, as illustrated in Figure 3.

After the mapping of a fully connected layer, we can obtain K_N query vectors $\mathbf{q}_j \in \mathbb{R}^d, 1 \leq j \leq K_N$. The key and value vectors $\mathbf{k} \in \mathbb{R}^{n \times d}$ are both obtained from the corresponding visual transformer layer's inputs, including image patch embeddings, image class token embedding, and visual prompts where n stands for their total numbers. Our proposed class-aware visual prompt $\tilde{\mathbf{P}}_l^j \in \mathbb{R}^d$ for the l_{th} layer is computed as

$$\mathbf{o}_l^j = \texttt{Softmax}(\frac{\mathbf{q}_j \mathbf{W}_q (\mathbf{k}\mathbf{W}_k)^T}{\sqrt{d_k}}) \mathbf{k}\mathbf{W}_v, 1 \leq j \leq K_N, \tag{6}$$

$$\tilde{\mathbf{P}}_l^j = LN(\mathbf{o}_l^j + \mathbf{q}_j), 1 \leq j \leq K_N, \tag{7}$$

where $LN(\cdot)$ denotes layer normalization. $\mathbf{W_q} \in \mathbb{R}^{d \times d_k}$, $\mathbf{W_k} \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W_v} \in \mathbb{R}^{d \times d}$ denote the parameters of cross attention.

To ensure the effect of the class-aware visual prompt, we additionally introduce a K-way classifier on top of the K_N outputs of the LN layer, and cross entropy loss is enforced on the K-way logits as follows:

$$\mathcal{L}_{ce}^{ca} = -\sum_i \mathbf{y}_i \log p_i, 1 \leq i \leq K, \tag{8}$$

where p_i denotes the i_{th} logit from classifying $LN(\mathbf{o}_l^j)$, K denotes the number of classes and $\mathbf{y}$ denotes the one-hot coding for the ground-truth target class. Note that only $\mathbf{o}_l^j$ derived from $\mathbf{q}_j$, which corresponds to the ground-truth target class, will be classified.

As the image class token embedding in deeper layers usually contains more task-related semantic information, the class-aware visual prompt is only applied to the last few layers of the image encoder in our implementation.

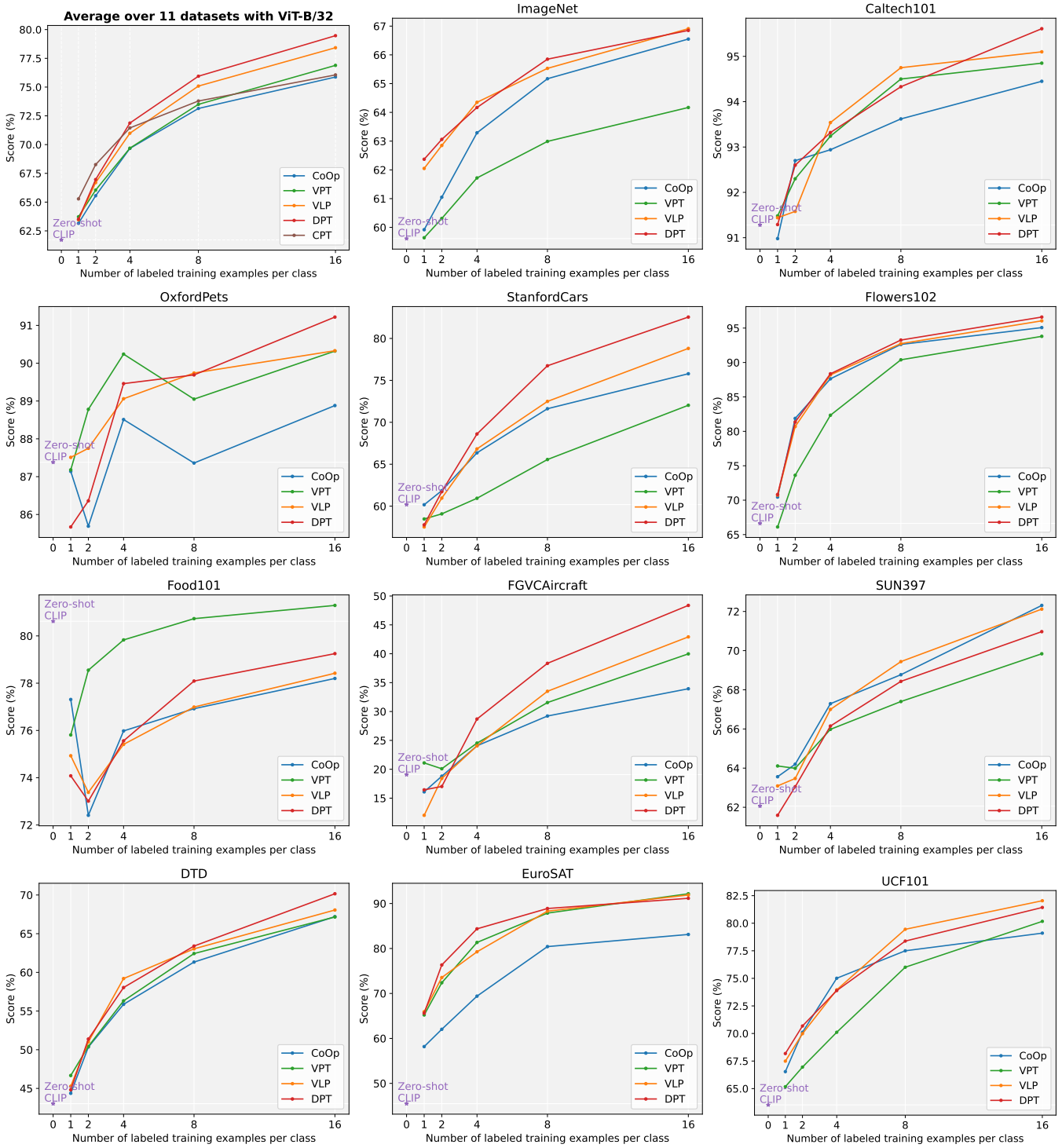

Fig. 4. Main results on the 11 datasets with 1,2,4,8,16 shots with ViT-B/32. Note that we also compare our methods with CPT [33] on average accuracy.

D. Training of DPT

A cross entropy loss is adopted to minimize the distance between the ground-truth annotation and the prediction probability computed by Equation (2).

$$\mathcal{L}_{ce} = -\sum_{i} \mathbf{y}_i \log p(y = i \mid \mathbf{x}''), 1 \leq i \leq K, \tag{9}$$

where $\mathbf{y}$ denotes the ground-truth annotation, $p(y = i \mid \mathbf{x}'')$ denotes the predicted probability from Equation (2), and $\mathbf{x}''$ is the final obtained image feature,

$$\mathbf{x}'' = f(\left[\mathbf{s}_0, \mathbf{P}_0, \cdots, \mathbf{P}_l, \tilde{\mathbf{P}}_{l+1}^{j}, \cdots, \tilde{\mathbf{P}}_{L}^{j}, \mathbf{E}_0\right]), \tag{10}$$

The total loss function combines the two cross-entropy losses with a balancing hyperparameter α as follows:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{ce}^{ca} + \mathcal{L}_{ce}. \tag{11}$$

E. Warm-up Strategy

To accelerate the training process, we adopt a general knowledge-guided warmup strategy in the first few epochs of training. Considering that the CLIP model stores general knowledge, we train our model to learn from zero-shot CLIP. The loss function we used for the first few epochs can be described as follows:

$$\mathcal{L} = \mathcal{L}_{coop} + \mathcal{L}_{vpt} + \beta\mathcal{L}_{ce} + \alpha\mathcal{L}_{ce}^{ca} \tag{12}$$

where $\mathcal{L}_{coop}$ is the loss function we used for CoOp training, $\mathcal{L}_{vpt}$ is the loss function we used in VPT training, and $\mathcal{L}_{ce}$ is the loss function we used in VLP training. β is a balancing hyperparameter. For $\mathcal{L}_{coop}$, we use the cross entropy loss to minimize the distance between the ground-truth annotation and the prediction probability computed by Equation (2).

$$\mathcal{L}_{coop} = -\sum_{i} \mathbf{y}_i \log p(y = i \mid \mathbf{x}), 1 \leq i \leq K, \tag{13}$$

For $\mathcal{L}_{vpt}$, the predicted probability is computed by Equation (1) instead of Equation (2).

$$\mathcal{L}_{vpt} = -\sum_{i} \mathbf{y}_i \log p(y = i \mid \mathbf{x}''), 1 \leq i \leq K, \tag{14}$$

By changing the loss function L_{ce} in the first few epochs of training to Equation (12), we use general knowledge to guide the warm-up process. During training, the proposed DPT keeps the entire parameters of both the image and text encoder fixed while optimizing the Text prompt, Visual prompt and the parameters for generating class-aware visual prompt.

F. Discussion on CAVPT

Fig. 3 illustrates the detailed computation process of the proposed class-aware visual prompts. As shown in Fig. 3, the CAVPT generator takes two types of inputs. Text prompt features from the text side include task-related information, while image patch embeddings from the image side represent visual instance information. First, the CAVPT generator performs cross-attention between the text prompt features and image patch embeddings, where query vectors are mapped from text prompt features while keys and values are derived from the image patch embeddings. Through the cross-attention operation, those image patch embeddings including more semantic information on objects belonging to the classes of downstream tasks will be more highlighted. As a result, the outputs of the cross-attention will include more features of the ground-truth objects. Then, our class-aware visual prompts are further generated with an additional "Add and Norm" operation similar to a typical transformer layer. As our class-aware visual prompts include richer semantic features of the ground-truth target objects, the final obtained image feature, which is computed by absorbing the information from the image patch embeddings and our class-aware visual prompts, can concentrate more on the classes corresponding to the downstream tasks.

IV. Experiments

A. Datasets and Implementation Details.

To evaluate the effectiveness of our method, we conducted experiments on 11 classification datasets, namely, EuroSAT [53], Caltech101 [54], OxfordFlowers [55], Food101 [56], FGVCAircraft [57], DTD [58], OxfordPets [59], StanfordCars [60], ImageNet1K [61], Sun397 [62], and UCF101 [63], as in [1], [9]. These datasets cover a wide range of computer vision tasks, including image classification on generic objects, fine-grained categories, satellite, texture, scene understanding and action recognition images.

Following the commonly used few-shot evaluation protocols in CLIP [1], we also adopted 1, 2, 4, 8, and 16 shots for model training and tested them on the full test dataset. The reported results are averaged over three runs for fair comparison.

We adopted ViT-Base/32 as our backbone network for all experiments. All experiments were conducted based on the official released code of CoOp [9] and CLIP [1] official released code. For VPT, the prompt length was set to 10 for each layer of the network, and they were initialized the same way as text prompts in CoOp [9]. During model training, we adopted the SGD optimization method, and the learning rate was decayed by the cosine rule. The maximum epoch for VPT was the same as CoOp [9]. The warm-up technique was adopted during the first 10 epochs with a fixed learning rate of 10^{-5} on VPT. The learning rate for VPT was first searched in $\{0.01, 0.001, 0.0001, 0.00001\}$ and kept unchanged for visual prompts in all experiments. For the text prompts, we followed CoOp [9] with a context length of 16.

For our proposed VLP and DPT methods, the maximum number of epochs was set to 100 for 16/8/4/2 shots, 60 epochs for 1 shot (the maximum number of epochs is set to 20 for ImageNet.) except for Caltech101 and OxfordPets in DPT, which was set to 60 for the 16-shot scenario. The warmup technique was the same as in CoOp and VPT (the warmup epoch was set to 1 for ImageNet on both ends). K_N was set to 10. CAVPT was inserted into the last layer of the image encoder.

The balancing α was set to 0.3, and β was set to 0.1. For the early training of VLP and DPT, general knowledge was utilized as guidance for the first 30 epochs (10 for ImageNet).

B. Comparison with Existing Methods

Existing representative prompt tuning methods include the remarkable CoOp method [9], and the CLIP model [1] itself used for zero-shot classification (i.e. Zero-Shot Clip). Therefore, we adopted these two models as our main comparison methods.

Since our DPT additionally introduces visual prompts and class-aware visual prompts compared with text prompts alone in CoOp, to reveal how each ingredient contributes to the performance improvements, we additionally implemented **VPT** and **VLP** in addition to **DPT** as follows:

• **VPT** standards for introducing a naive visual prompt alone into the visual end of the CLIP model [1] and hand-crafted

TABLE II
RESULTS OF 11 DATASETS UNDER 16-SHOTS SETTING WITH VIT-B/16.

Methods	EuroSAT	Caltech101	Oxford Flowers	Food101	FGVC Aircraft	DTD	OxfordPets	Stanford Cars	Sun397	UCF101	ImageNet	Average
ZSCLIP [1]	47.69	93.75	70.69	85.97	24.81	43.09	89.07	65.55	62.61	67.54	64.51	65.03
CoOp [9]	83.74	95.17	96.73	84.17	44.06	69.60	92.07	82.73	74.54	82.59	71.62	79.73
CoCoOp [15]	72.07	95.71	88.74	**87.37**	30.09	62.53	93.33	71.60	72.36	77.90	70.38	74.73
ProGrad [26]	84.29	95.89	96.30	86.68	41.23	68.83	93.25	81.71	75.10	81.16	71.94	79.67
ProDA [27]	85.17	96.23	97.54	87.29	44.40	72.46	93.42	83.89	**77.19**	85.12	**72.73**	81.40
VPT	**92.67**	**96.27**	96.59	87.03	51.11	71.26	92.76	81.44	72.93	85.19	69.98	81.57
VLP	91.87	96.08	97.37	84.57	52.99	72.20	93.11	85.62	74.48	**86.36**	72.46	82.46
DPT	92.10	96.06	**97.59**	85.00	**57.85**	**72.65**	**93.45**	**88.24**	74.29	85.31	72.49	**83.18**

text prompts, *e.g.* "a photo of a [CLASS]", were adopted for the text end.

• **VLP** denotes the dual modality prompt tuning paradigm to simultaneously learn visual (V) and text (L) prompts, where the text prompt was designed in the same way as that in CoOp [9], and the visual prompt was exactly the same as VPT-Deep in VPT [23].

• **DPT** indicates that we further integrated CAVPT into the image encoder based on VLP.

The overall evaluation results are shown in Figure 4, which reports the classification accuracy on 11 datasets under all few-shot settings. Compared with the baseline methods, our DPT achieved superior performances on average over the 11 datasets. Figure 4 and Table I clearly show the following: 1) DPT greatly outperforms CoOp and zero-shot CLIP by large margins. The performance increases are basically proportional to the number of shots. Specifically, DPT outperforms zero-shot CLIP by 17.6% and outperforms CoOp by 3.53% on average over the 11 datasets under the 16-shot settings. The results verified the superiority of the proposed DPT paradigm. 2) Comparing the results of VPT with CoOp, VPT obtains better results than CoOp on average. Under the 16-shot setting, VPT can obtain 1% performance gains on average over all 11 datasets. It shows that tuning the visual prompts from the image end instead of text prompts can obtain more effective results. It is worth noting that VPT and CoOp obtain inconsistent results on different datasets, which indicates that tuning visual prompts and text prompts may have complementary effects. 3) Comparing the results of VLP with those of VPT and CoOp, VLP achieves better results than VPT and CoOp, which shows that tuning the dual modality prompts from both the visual and text ends is obviously better than tuning any single modality prompts for the downstream task. 4) With the help of class-aware visual prompts, the results of our DPT are clearly improved compared to VLP. Specifically, DPT obtains 1% performance gains over VLP on average over 11 datasets under a 16-shot setting, which shows the great significance of our CAVPT.

C. Domain Generalization

In this section, we aim to unveil how robust our method is to distribution shifts in comparison to baseline methods.

Datasets. Following the setting in CoOp [9], we used ImageNet as the source dataset. The target datasets were ImageNetV2 [64], ImageNet-Sketch [65], ImageNet-A [66], and ImageNet-R [67].

Setting. We chose CoOp and VPT as our baseline methods. All three methods were trained on the source dataset with 1 example per class, and zero-shot inference was conducted on the target datasets.

Results. As shown in Table III, our method achieves the best performances on ImageNet, ImageNetV2, ImageNet-Sketch, and ImageNet-R, while our method is less effective on ImageNet-A, which contains natural adversarial examples. This suggests that our method has stronger robustness than baseline methods but tends to be more vulnerable when facing adversarial examples compared with CoOp. In contrast, the VPT model obtains inferior results on target datasets, which shows that VPT is less robust than CoOp and our method.

D. Further Analysis

Analysis of the depth of CAVPT insertion. CAVPT is a plug-and-play module and can be used in arbitrary layers of the ViT backbone. To investigate the most suitable layers for CAVPT, we conducted comprehensive experiments in both bottom-up and top-down fashions with varying values of depth, *i.e.* $\{1 \rightarrow 12, 4 \rightarrow 12, 8 \rightarrow 12, 12\}$ for top-down fashion and $\{1, 1 \rightarrow 4, 1 \rightarrow 8, 1 \rightarrow 12\}$ for bottom-up fashion, on top of the VLP model, on all 11 datasets. An extra experimental setting of sharing CAVPT across different layers was also conducted in the top-down fashion by sharing the parameters of CAVPT. As shown in Figure 6, the results of the top-down fashion are much better than those of the bottom-up fashion, and the last layer of the Transformer is the most suitable layer for CAVPT, suggesting that CAVPT plays a more important role at deeper layers. Additionally, comparing the shared CAVPT and vanilla CAVPT, the shared CAVPT can achieve slightly better results while having fewer parameters.

Analysis of the length of CAVPT. To investigate the suitable length of CAVPT, we conducted comprehensive experiments on different lengths of CAVPT, *i.e.* $\{0, 1, 5, 10, 20, 50, 100\}$. A length of 0 indicates that the method deteriorates to VLP but without visual prompts in the last layer of the ViT backbone. For some datasets, taking EuroSAT [53] as an example, it only contains 10 classes, which is insufficient to obtain more than 10 CAVPTs. Thus, when the number of required CAVPTs is larger than the

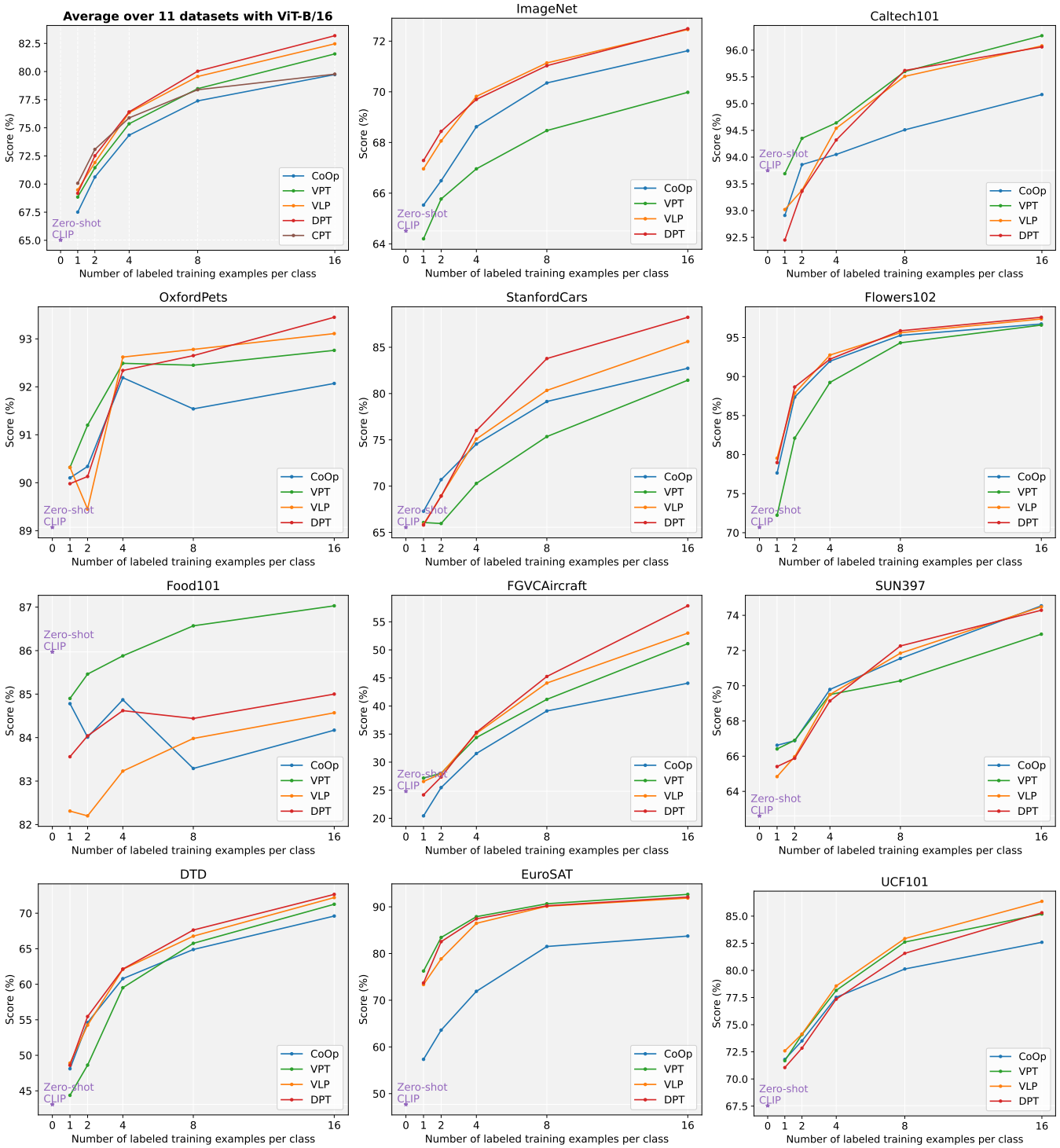

Fig. 5. Main results on the 11 datasets with 1,2,4,8,16 shots with ViT-B/16. Note that we also compare our methods with CPT [33] on average accuracy.

number of classes, we take the number of classes as the length of the CAVPT to obtain the corresponding results. As shown in Table V, setting the length to 10 achieves the best accuracy.

Analysis of the loss function on the CAVPT module. In the proposed CAVPT module, we apply the cross-entropy loss to encourage the alignment of the visual class token and text prompt features. To demonstrate the effectiveness of the loss function, we optimized the model with different α on the CAVPT module. The experimental results are shown in Table IV. We can clearly see that setting $\alpha = 0.3$ helps to improve the average accuracy by 0.44%, which significantly illustrates the efficacy of such a loss function.

Analysis of the parameters of different models. Since DPT introduces more parameters than VLP and VPT, the question arises: can VLP or VPT achieve the same performance as DPT with the same number of parameters? We increased

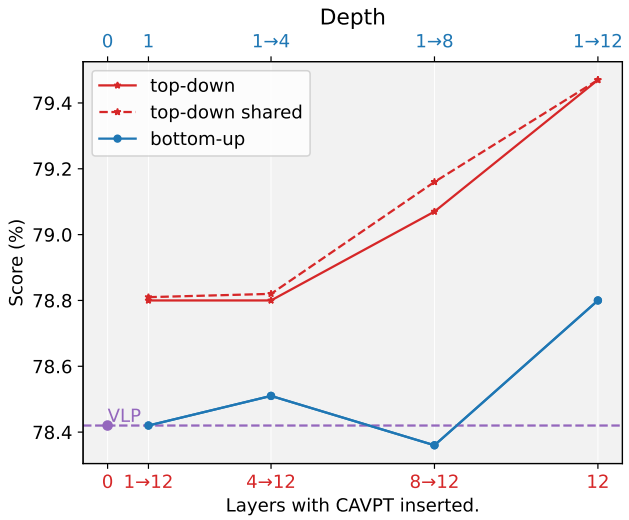

Fig. 6. The average accuracy on 11 datasets with different layers CAVPT inserted. $i \rightarrow j$ indicates the transformer layer into which CAVPT is inserted.

the number of visual prompts to 120 for VLP and VPT to compare their performance with DPT under a competitive number of parameters. Note that CoOp limits the input token number. Thus, CoOp is not discussed in this section. As shown in Table VI, with a larger number of visual prompts, both VPT and VLP performance dropped drastically, suggesting that simply enlarging the parameter size will hamper the performance.

TABLE III
COMPARISION WITH SINGLE MODAL PROMPT METHODS ON ROBUSTNESS TO DISTRIBUTION SHIFT UNDER 1-SHOT SCENARIO.

Method	ImageNet	-V2	-S	-A	-R	Average	OOD Average
CoOp [9]	59.92	52.88	37.32	**28.52**	62.12	48.15	45.21
VPT	59.64	52.18	35.74	21.31	59.93	45.76	42.29
DPT	**62.37**	**55.15**	**39.65**	27.79	**64.79**	**49.95**	**46.85**

TABLE IV
THE AVERAGE ACCURACY ON 11 DATASETS WITH DIFFERENT α.

α	0	0.1	0.3	0.5	0.7	1
Average	78.96	79.27	**79.47**	79.25	79.28	79.27

TABLE V
THE AVERAGE ACCURACY ON 11 DATASETS WITH VARIOUS LENGTHS OF CAVPT.

Length	0	1	5	10	20	50	100
Average	78.41	79.21	79.29	**79.47**	79.35	79.28	79.33

Analysis of different backbones. To further show the effectiveness of our method, we conducted experiments on the ViT-B/16 backbone. As shown in Fig 5 and Table II, DPT outperforms zero-shot CLIP and CoOp by 18.15% and 3.45% on average over the 11 datasets under 16-shot settings, which demonstrates the superiority of DPT with other backbones.

TABLE VI
DPT VS BIGGER VLP VS VPT ON 11 DATASETS UNDER 16 SHOTS SETTING. VCTX STANDS FOR THE NUMBER OF VISUAL PROMPTS.

Methods	# params	Average
VPT(VCTX=10)	92,160	76.89
VPT(VCTX=120)	1,105,920	73.64
VLP(VCTX=10)	100,352	78.42
VLP(VCTX=120)	1,114,112	71.71
DPT	1,136,384	**79.47**

The same conclusions can also be drawn that tuning the visual prompts is more effective than text prompts, and the joint tuning of visual-text prompts also boosts the classification accuracy.

In summary, the experimental results under the ViT-B/16 backbone are consistent with those under the ViT-B/32 backbone, which indicates the effectiveness and reasonability of dual modality prompt tuning.

E. Visualization of the attention map.

In Figure 1, we visualize and compare the attention map for the last layer of CoOp and our DPT tuned model to understand the proposed method in depth. Figure 1 (a) shows the original images with target objects in red bounding boxes. Figure 1 (b) shows the attention maps of the baseline method Zero-shot CLIP/CoOp. As CoOp does not tune the image encoder, the attention maps are the same with zero-shot CLIP. Figure 1 (c) depicts the attention maps of our proposed DPT method. It can be clearly seen that Zero-shot CLIP/CoOp usually focuses on most of the typical objects in the image, while DPT tends to concentrate more on the target visual object (concept).

We show extra examples of visualization in Fig 7. All of these examples are sampled from Caltech101, StanfordCars, Food101, FGVCAircraft and OxfordPets. In Figure 7(a), we annotated the object of interest in the red box. Figure 7(b) demonstrates the attention map of Zeroshot CLIP/CoOp. As CoOp has no modification on image features, the attention maps are the same with ZS CLIP. It can be clearly seen that multiple objects are highlighted while only a little attention the model pays to the objects of interest in downstream tasks. In Figure 7(c), which shows the visualization of VPT, the object of interest is well highlighted. This shows that VPTs learned some downstream task-related knowledge. The visualization of VLP and DPT are shown in Figure 7(d) and (e). Comparing Figure 7(d) and Figure 7(e), the object of interest would be more concentrated, and more non-related objects are less highlighted in Figure 7(e). It shows that CAVPT can help the model to pay more attention to the right object rather than other task unrelated objects.

The last two rows of Fig. 7 show some typical failure cases. It can be clearly seen that the regions corresponding to the ground-truth target classes on the visualized attention maps for ZSCLIP and VPT are not highlighted, i.e., The visual features corresponding to the target classes are not significant in the oracle image features extracted from the pretrained foundation models. As analyzed above, text prompts can serve as synthesized classifiers, while visual prompts are expected

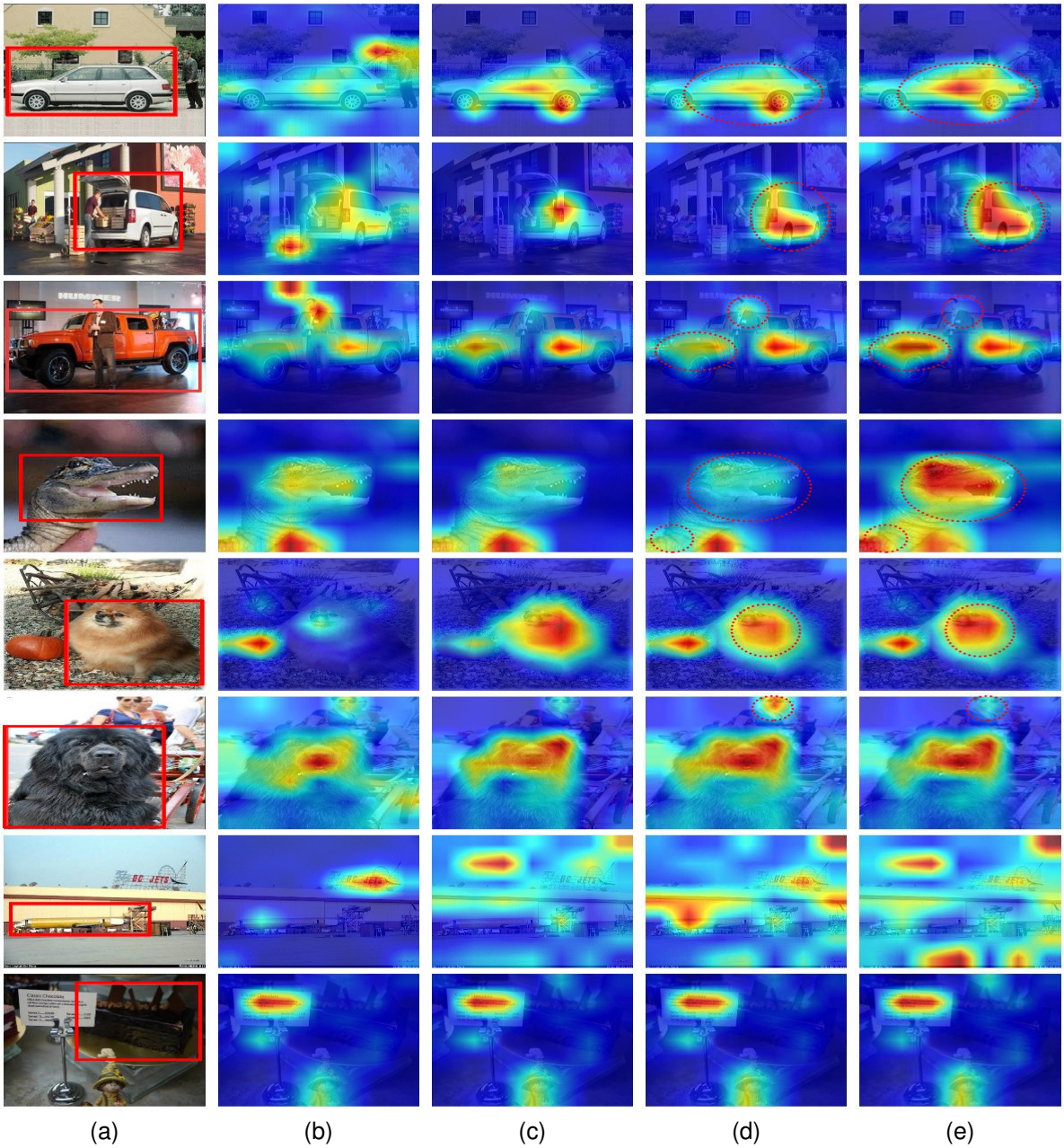

Fig. 7. Comparison of attention map visualization of the variant methods. (a) Original Image. (b) Zero-shot CLIP/CoOp. (c) VPT-Deep. (d) VLP. (e) DPT. The GT annotated object is marked by a red box. The last two rows are failure cases where the model fails to focus on the GT annotated object.

to query suitable knowledge stored in the pretrained image encoder. If the visual features are not very significant in the pretrained foundation image encoder, it is not easy to adjust the obtained image features to focus on the target class, especially in the few-shot cases.

V. Conclusion

In this paper, we propose a new dual-modality prompt tuning paradigm for tuning the large pretrained vision-language model to downstream tasks by learning the visual and text prompts simultaneously. To make the final obtained image feature concentrate more on the target visual concept, we further encode both the downstream task-related information and image instance information into the visual prompt and propose class-aware visual prompts, which are dynamically generated by performing cross attention between text prompt features and image token embeddings. Extensive experimental results on 11 datasets demonstrate the effectiveness of the proposed method and show its superiority to other prompt tuning approaches by a large margin.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[3] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 162–11 173.

[4] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4222–4235.

[7] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[9] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, pp. 1–12, 2022.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[11] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[12] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," *arXiv preprint arXiv:2110.05208*, 2021.

[13] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[14] R. Zhang, R. Fang, P. Gao, W. Zhang, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.

[15] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[16] H. Fang, P. Xiong, L. Xu, and W. Luo, "Transferring image-clip to video-text retrieval via temporal relations," *IEEE Transactions on Multimedia*, 2022.

[17] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for multi-modal tracking," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3492–3500.

[18] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 082–18 091.

[19] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?" in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6868–6874.

[20] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[21] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562.

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[23] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," *arXiv preprint arXiv:2203.12119*, 2022.

[24] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.

[25] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, 2022.

[26] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," *arXiv preprint arXiv:2205.14865*, 2022.

[27] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5206–5215.

[28] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, vol. 1, no. 3, p. 4, 2022.

[29] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," *arXiv preprint arXiv:2206.04673*, 2022.

[30] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 1–18.

[31] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[32] H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, and H. Yang, "Prompt tuning for generative multimodal pretrained models," *arXiv preprint arXiv:2208.02532*, 2022.

[33] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "Cpt: Colorful prompt tuning for pre-trained vision-language models," *arXiv preprint arXiv:2109.11797*, 2021.

[34] A. Sahoo, A. Senapati, A. Das, Y. Kim, R. Feris, and R. Panda, "Frustratingly simple contrastive prompt tuning for vision-language models."

[35] Y. Lu, W. Wang, C. Yuan, X. Li, and Z. Lai, "Manifold transfer learning via discriminant regression analysis," *IEEE Transactions on Multimedia*, vol. 23, pp. 2056–2070, 2020.

[36] P. Jing, Y. Su, L. Nie, and H. Gu, "Predicting image memorability through adaptive transfer learning from external sources," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1050–1062, 2016.

[37] J. O. Zhang, A. Sax, A. Zamir, L. Guibas, and J. Malik, "Side-tuning: a baseline for network adaptation via additive side networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 2020, pp. 698–714.

[38] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce memory, not parameters for efficient on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 285–11 297, 2020.

[39] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[40] Y. Zhu, W. Min, and S. Jiang, "Attribute-guided feature learning for few-shot image recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1200–1209, 2020.

[41] H. Zhang, H. Li, and P. Koniusz, "Multi-level second-order few-shot learning," *IEEE Transactions on Multimedia*, 2022.

[42] H. Cheng, J. T. Zhou, W. P. Tay, and B. Wen, "Graph neural networks with triple attention for few-shot learning," *IEEE Transactions on Multimedia*, 2023.

[43] K. Guo, C. Shen, B. Hu, M. Hu, and X. Kui, "Rsnet: relation separation network for few-shot similar class recognition," *IEEE Transactions on Multimedia*, 2022.

[44] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2020.

[45] P. Tian, H. Yu, and S. Xie, "An adversarial meta-training framework for cross-domain few-shot learning," *IEEE Transactions on Multimedia*, 2022.

[46] X. Zhong, C. Gu, M. Ye, W. Huang, and C.-W. Lin, "Graph complemented latent representation for few-shot image classification," *IEEE Transactions on Multimedia*, 2022.

[47] L. Zhang, Y. Du, J. Shen, and X. Zhen, "Learning to learn with variational inference for cross-domain image classification," *IEEE Transactions on Multimedia*, 2022.

[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[49] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning,"

Advances in neural information processing systems, vol. 33, pp. 21 271–21 284, 2020.

[50] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[53] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[54] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[55] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[56] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101–mining discriminative components with random forests," in *European conference on computer vision*. Springer, 2014, pp. 446–461.

[57] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[58] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[59] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[60] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[62] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[63] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[64] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.

[65] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[66] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[67] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.